# Screening for Chronic Kidney Disease

Rashi Desai

663553314

# Screening for Chronic Kidney Disease

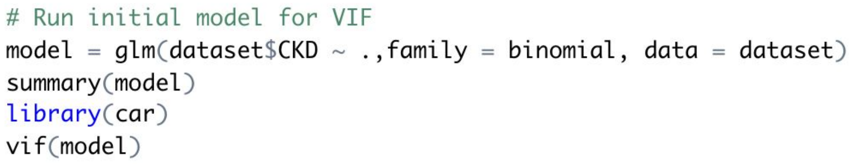
- CDC and NCHS collects data from nationwide surveys of US adults
- Here, we have a dataset of 8819 adults:
  6000 records of training data and predict 2819 records of test data
- Problem Statement
  Identify patients at risk of having Chronic Kidney Disease from a dataset with 34 variables to get tested in case of high probability
- Dependent variable: CKD - 34th Variable (a binomial variable [1,0])
- Dataset consists of 10 continuous variables and 23 categorical variables

# Data Pre-Processing

- Removed 2819 rows with no prediction if the patient has CKD for training dataset
- Removed rows with missing values in the dataset rather than imputing the data
- Remove highly correlated variables using VIF
- Use the remaining important predictor variables to run a Logistic Regression model on training data
- Use the same model to predict CKD in test data

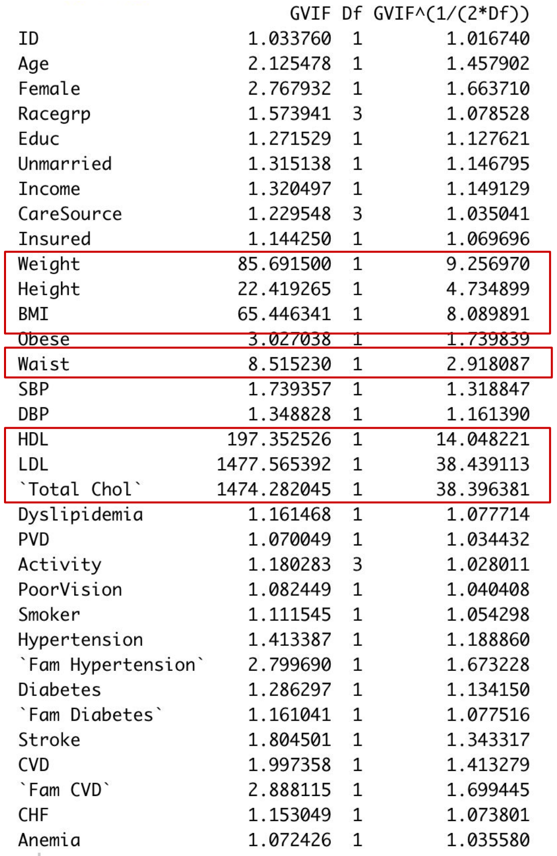
# Exploratory Data Analysis

Removed multi-collinearity among predictors by VIF (variance inflation factor) and correlation to filter variables



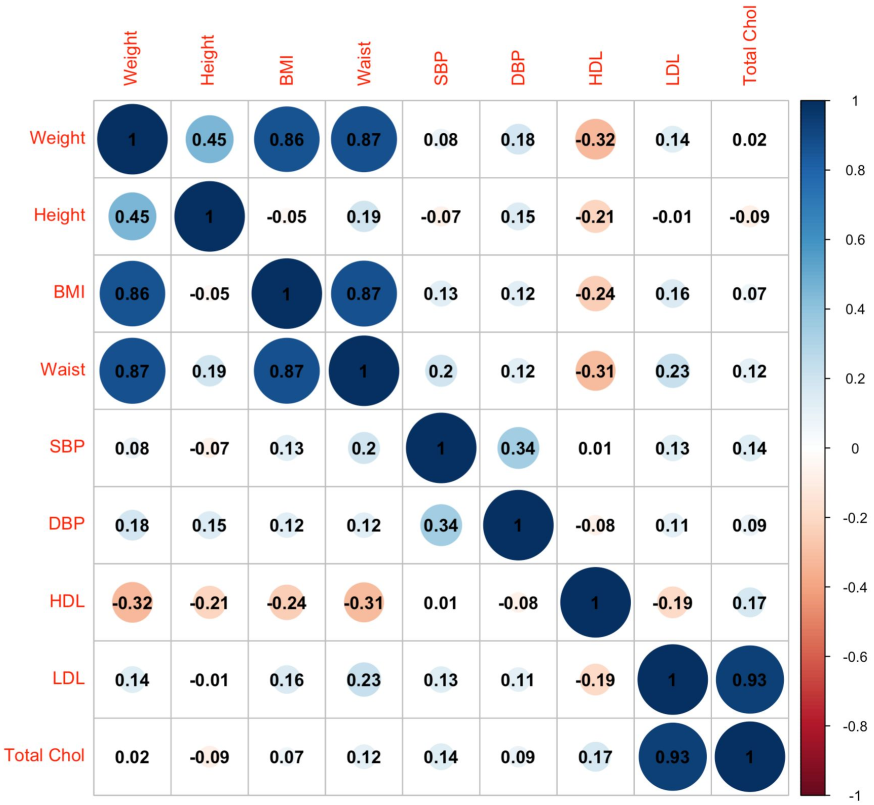
```
# Run initial model for VIF
model = glm(dataset$CKD ~ .,family = binomial, data = dataset)
summary(model)
library(car)
vif(model)
```

| Removed | High Correlation with |
|---|---|
| Height, Obese, Waist & BMI | Weight |
| Total Chol | LDL |
| Fam Hypertension | Hypertension |
| Fam Diabetes | Diabetes |
| Fam CVD | CVD |



```
                        GVIF Df GVIF^(1/(2*Df))
ID                  1.033760  1        1.016740
Age                 2.125478  1        1.457902
Female              2.767932  1        1.663710
Racegrp             1.573941  3        1.078528
Educ                1.271529  1        1.127621
Unmarried           1.315138  1        1.146795
Income              1.320497  1        1.149129
CareSource          1.229548  3        1.035041
Insured             1.144250  1        1.069696
Weight             85.691500  1        9.256970
Height             22.419265  1        4.734899
BMI                65.446341  1        8.089891
Obese               3.027038  1        1.739839
Waist               8.515230  1        2.918087
SBP                 1.739357  1        1.318847
DBP                 1.348828  1        1.161390
HDL               197.352526  1       14.048221
LDL              1477.565392  1       38.439113
`Total Chol`     1474.282045  1       38.396381
Dyslipidemia        1.161468  1        1.077714
PVD                 1.070049  1        1.034432
Activity            1.180283  3        1.028011
PoorVision          1.082449  1        1.040408
Smoker              1.111545  1        1.054298
Hypertension        1.413387  1        1.188860
`Fam Hypertension`  2.799690  1        1.673228
Diabetes            1.286297  1        1.134150
`Fam Diabetes`      1.161041  1        1.077516
Stroke              1.804501  1        1.343317
CVD                 1.997358  1        1.413279
`Fam CVD`           2.888115  1        1.699445
CHF                 1.153049  1        1.073801
Anemia              1.072426  1        1.035580
```

# Exploratory Data Analysis



| | Weight | Height | BMI | Waist | SBP | DBP | HDL | LDL | Total Chol |
|---|---|---|---|---|---|---|---|---|---|
| Weight | 1 | 0.45 | 0.86 | 0.87 | 0.08 | 0.18 | -0.32 | 0.14 | 0.02 |
| Height | 0.45 | 1 | -0.05 | 0.19 | -0.07 | 0.15 | -0.21 | -0.01 | -0.09 |
| BMI | 0.86 | -0.05 | 1 | 0.87 | 0.13 | 0.12 | -0.24 | 0.16 | 0.07 |
| Waist | 0.87 | 0.19 | 0.87 | 1 | 0.2 | 0.12 | -0.31 | 0.23 | 0.12 |
| SBP | 0.08 | -0.07 | 0.13 | 0.2 | 1 | 0.34 | 0.01 | 0.13 | 0.14 |
| DBP | 0.18 | 0.15 | 0.12 | 0.12 | 0.34 | 1 | -0.08 | 0.11 | 0.09 |
| HDL | -0.32 | -0.21 | -0.24 | -0.31 | 0.01 | -0.08 | 1 | -0.19 | 0.17 |
| LDL | 0.14 | -0.01 | 0.16 | 0.23 | 0.13 | 0.11 | -0.19 | 1 | 0.93 |
| Total Chol | 0.02 | -0.09 | 0.07 | 0.12 | 0.14 | 0.09 | 0.17 | 0.93 | 1 |

Two Sample t-test between continuous and target variables to include only strong predictors

| Variable 1 | Variable 2 | P-value |
|---|---|---|
| Weight | CKD | 0.7206 |
| Height | CKD | 0.02718 |
| BMI | CKD | 0.1495 |
| Waist | CKD | 7.242e-07 |
| SBP | CKD | < 2.2e-16 |
| DBP | CKD | 0.007493 |
| HDL | CKD | 0.002158 |
| LDL | CKD | 0.02037 |
| Total Chol | CKD | 0.1795 |

# Exploratory Data Analysis

Chi-square test of all categorical variables with the target variable 'CKD' to find the significant variables

| Variable 1 | Variable 2 | P-value |
|---|---|---|
| Female | CKD | 0.5182 |
| Education | CKD | 4.349e-05 |
| Unmarried | CKD | 0.00145 |
| Income | CKD | 3.491e-08 |
| CareSource | CKD | 4.259e-07 |
| Insured | CKD | 3.439e-11 |
| Obese | CKD | 0.1664 |
| Dyslipidemia | CKD | 1.00 |
| PVD | CKD | < 2.2e-16 |
| Activity | CKD | 1.29e-09 |

| Variable 1 | Variable 2 | P-value |
|---|---|---|
| Poor Vision | CKD | 3.39e-10 |
| Smoker | CKD | < 2.2e-16 |
| Stroke | CKD | < 2.2e-16 |
| CVD | CKD | < 2.2e-16 |
| CHF | CKD | 1.032e-12 |
| Anemia | CKD | 0.2474 |

# Data Modelling

Removing the columns that were not found significant [21]:

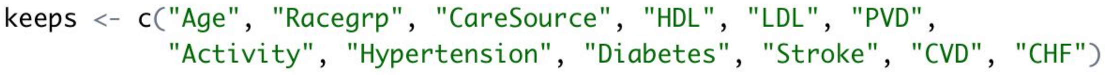
ID, Educ, Unmarried, Income, Insured, Weight, Height, BMI, Obese, Waist, Total Chol, SBP, DBP, Dyslipidemia, Poor Vision, Smoker, Fam Hypertension, Fam Diabetes, Fam CVD, Anemia

Variables that remain for data modelling [12]:

Age, Female, Racegrp, CareSource, HDL, LDL, PVD, Activity, Hypertension, Diabetes, Stroke, CVD, CHF

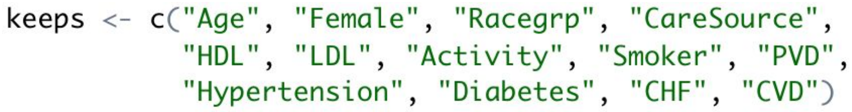
# Exploratory Data Analysis

## Model 1



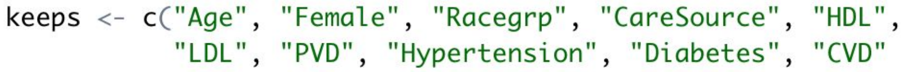
```
keeps <- c("Age", "Racegrp", "CareSource", "HDL", "LDL", "PVD",
           "Activity", "Hypertension", "Diabetes", "Stroke", "CVD", "CHF")
```

## Model 2



```
keeps <- c("Age", "Female", "Racegrp", "CareSource",
           "HDL", "LDL", "Activity", "Smoker", "PVD",
           "Hypertension", "Diabetes", "CHF", "CVD")
```

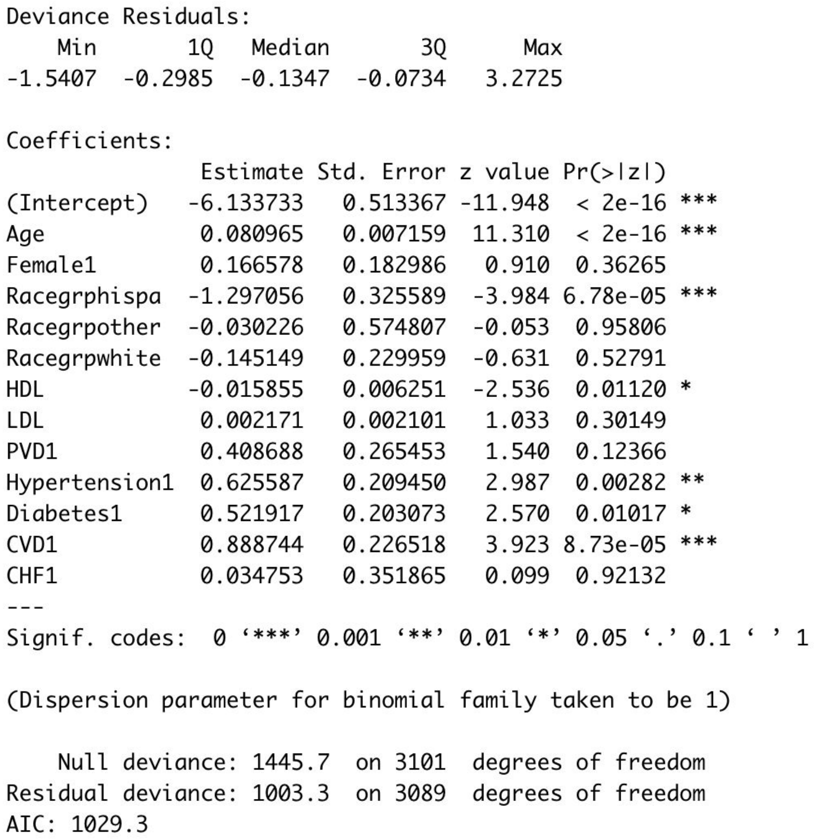
## Model 3



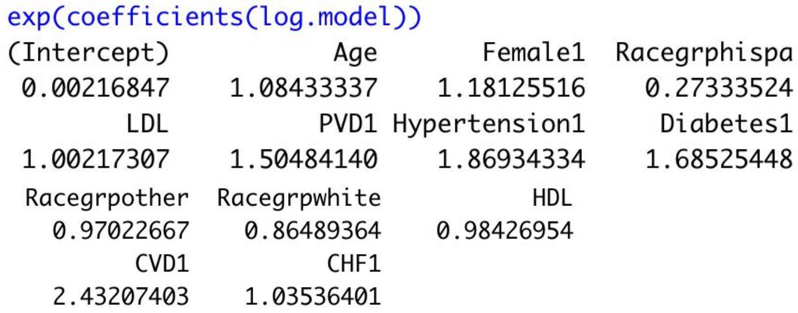
```
keeps <- c("Age", "Female", "Racegrp", "CareSource", "HDL",
           "LDL", "PVD", "Hypertension", "Diabetes", "CVD"
```

| Model | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| 1 | 75.0% | 87.1% | 86.3% |
| 2 | 80.8% | 84.58% | 84.3% |
| 3 | 82.4% | 85.0% | 84.7% |

# Data Modelling (Full Logistic Regression)



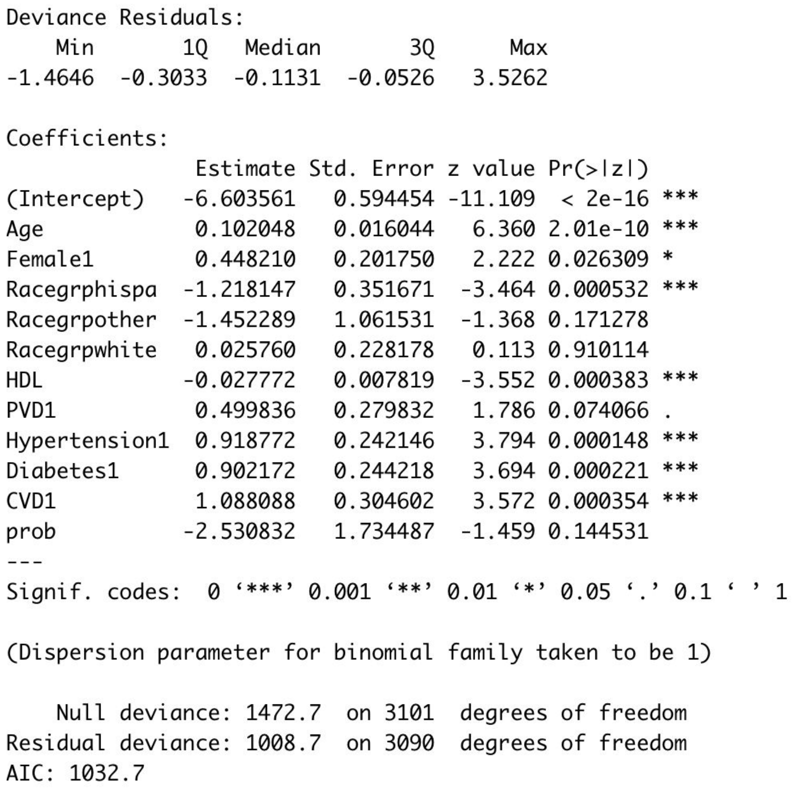
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5407  -0.2985  -0.1347  -0.0734   3.2725

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.133733   0.513367 -11.948  < 2e-16 ***
Age           0.080965   0.007159  11.310  < 2e-16 ***
Female1       0.166578   0.182986   0.910  0.36265
Racegrphispa -1.297056   0.325589  -3.984 6.78e-05 ***
Racegrpother -0.030226   0.574807  -0.053  0.95806
Racegrpwhite -0.145149   0.229959  -0.631  0.52791
HDL          -0.015855   0.006251  -2.536  0.01120 *
LDL           0.002171   0.002101   1.033  0.30149
PVD1          0.408688   0.265453   1.540  0.12366
Hypertension1 0.625587   0.209450   2.987  0.00282 **
Diabetes1     0.521917   0.203073   2.570  0.01017 *
CVD1          0.888744   0.226518   3.923 8.73e-05 ***
CHF1          0.034753   0.351865   0.099  0.92132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1445.7  on 3101  degrees of freedom
Residual deviance: 1003.3  on 3089  degrees of freedom
AIC: 1029.3
```



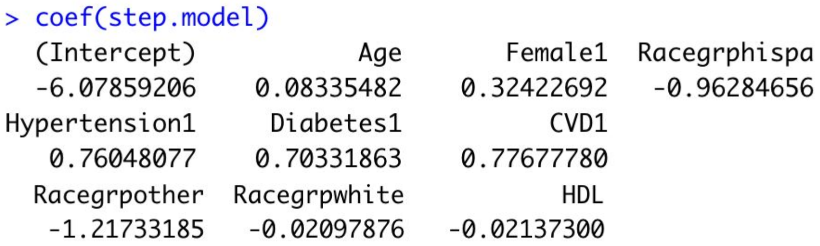
```
exp(coefficients(log.model))
 (Intercept)          Age      Female1  Racegrphispa
  0.00216847   1.08433337   1.18125516   0.27333524
         LDL         PVD1 Hypertension1     Diabetes1
  1.00217307   1.50484140   1.86934334   1.68525448
 Racegrpother Racegrpwhite          HDL
   0.97022667   0.86489364   0.98426954
         CVD1          CHF1
   2.43207403   1.03536401
```

| Model | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| Full | 82.4% | 85.9% | 85.57% |

# Data Modelling (Stepwise Logistic Regression)



```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4646  -0.3033  -0.1131  -0.0526   3.5262  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -6.603561   0.594454 -11.109  < 2e-16 ***
Age            0.102048   0.016044   6.360 2.01e-10 ***
Female1        0.448210   0.201750   2.222 0.026309 *  
Racegrphispa  -1.218147   0.351671  -3.464 0.000532 ***
Racegrpother  -1.452289   1.061531  -1.368 0.171278    
Racegrpwhite   0.025760   0.228178   0.113 0.910114    
HDL           -0.027772   0.007819  -3.552 0.000383 ***
PVD1           0.499836   0.279832   1.786 0.074066 .  
Hypertension1  0.918772   0.242146   3.794 0.000148 ***
Diabetes1      0.902172   0.244218   3.694 0.000221 ***
CVD1           1.088088   0.304602   3.572 0.000354 ***
prob          -2.530832   1.734487  -1.459 0.144531    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1472.7  on 3101  degrees of freedom
Residual deviance: 1008.7  on 3090  degrees of freedom
AIC: 1032.7
```

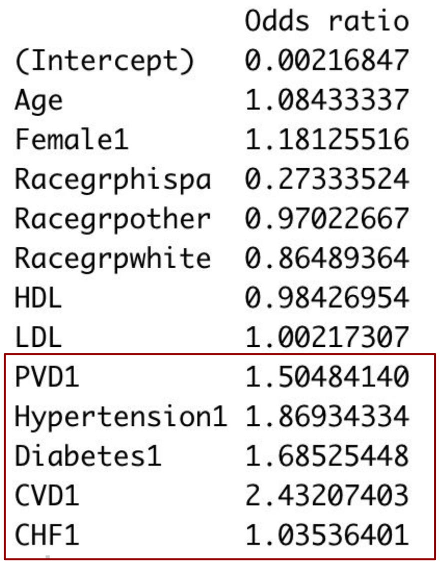


```
> coef(step.model)
  (Intercept)           Age       Female1  Racegrphispa 
  -6.07859206    0.08335482    0.32422692   -0.96284656 
Hypertension1     Diabetes1          CVD1 
   0.76048077    0.70331863    0.77677780 
 Racegrpother  Racegrpwhite           HDL 
  -1.21733185   -0.02097876   -0.02137300 
```

| Model | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| Stepwise | 76.4% | 86.1% | 85.4% |

# Odds Ratio

- Odds = $\frac{p\ (occurring)}{1-p\ (not\ occurring)}$
- Odds Ratio = $\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$
- Odds Ratio indicate how odds change with 1 unit increase in a variable holding other variable constants
- CVD increases the odds of having CKD by 143%
- Hypertension increases the odds of 'Chronic Kidney Disease' by 86.93%
- Diabetes increases the odds of 'Chronic Kidney Disease' by 68.5%



```
              Odds ratio
(Intercept)   0.00216847
Age           1.08433337
Female1       1.18125516
Racegrphispa  0.27333524
Racegrpother  0.97022667
Racegrpwhite  0.86489364
HDL           0.98426954
LDL           1.00217307
PVD1          1.50484140
Hypertension1 1.86934334
Diabetes1     1.68525448
CVD1          2.43207403
CHF1          1.03536401
```

# Results

From the two regression models and odd ratios, we can conclude that people have a higher risk of getting a CKD if they have:

  - Cardiovascular Diseases
  - Diabetes
  - Hypertension
  - High Levels of HDL
- Adults of the race Hispanic have a higher risk of CKD
- Adults of age greater than 60 have a higher risk of CKD